Metadatapedia: A proposal for aggregating metadata on data archiving

David M. Nichols
Department of Computer Science
University of Waikato
Hamilton, New Zealand
+64 7 8585130
dmn@cs.waikato.ac.nz

Michael B. Twidale
Graduate School of Library and Information Science
University of Illinois, USA
+ 1 217 265-0510
twidale@illinois.edu

Sally Jo Cunningham
Department of Computer Science
University of Waikato
Hamilton, New Zealand
+64 7 8384402
sallyjo@cs.waikato.ac.nz

ABSTRACT
The open access movement has highlighted the barriers that exist for users to gain access to significant portions of the research literature. The open data approach seeks to extend the principles of open access to the data and code that supports the published scholarly record. Current metadata is inadequate to allow information researchers to evaluate claims made about data archiving practices. Assessing current archiving practice and understanding the impact of archiving policies requires improved metadata. We propose that information researchers create an infrastructure for the collection of metadata about data use in the research literature, and that infrastructure should itself be open. The availability of metadata on data use would enable the calculation of archiving indices, just as citation data enables the calculation of the h-index.

Categories and Subject Descriptors
E. Data

General Terms
Documentation

Keywords
Metadata, data archiving, data curation, open data.

1. INTRODUCTION
The open access movement has highlighted the barriers that exist for users to gain access to significant portions of the research literature [30]. Although access to the published articles remains an important issue, access to the underlying data and code that supports the research findings is becoming a significant practical and policy issue [3,7,10,11,26,27].

Several funding agencies have established archiving requirements for the research papers derived from their grants, e.g. Science Foundation Ireland (SFI) policy states that "all researchers are required to lodge their publications resulting in whole or in part from SFI-funded research in an open access repository as soon as possible after publication" [23].

Archiving policies are now being extended to the data that may be associated with funding grants; the NSF Data Management Plan [19] is a high-profile example. Some journals are also requiring authors to provide information on where and how supporting data may be obtained [16]. It seems likely that we are seeing the start of a long-term trend of increasing the quantity of data archives associated with research papers. This raises questions of verification: how easy is it to check if a data policy is being complied with? More interestingly, we can expect to see more measurement of metadata related to the impact of data archiving policies, just as we already see studies relating to the open access archiving of papers (e.g. [31]).

The adequacy, or otherwise, of research data archiving can be brought into sharp relief by considering communities who are critical of current practice. One of these communities can be loosely termed the 'climate sceptic' blogosphere; although covering a wide range of opinions, a common theme is that the consensus results of climate science are either not correct and/or not reproducible. An interesting feature of this community is its willingness to engage with primary research. One of the many themes that emerge in their online discussions is the adequacy of several aspects of data archiving in the research literature on climate; here is one example:

> *Lonnie Thompson is one of the worst archiving offenders in paleoclimate, and that's a real beauty contest.* [18]

This statement identifies one researcher as being among those with the lowest rates of data archiving within one research community. It further implicitly claims that this particular community, "paleoclimate", is on average worse at data archiving than other communities. This is clearly a provocative assertion, and intended to be so. From an information research perspective it seems reasonable that we should be able to assess the truth of such claims; however, we currently cannot tell whether this claim is true or false. Furthermore, without a change in the metadata environment, there is little likelihood that we will ever be able to confirm or refute similar claims in any discipline.

The metadata needed to assess such claims is simply *missing*. Although some work [e.g. 1,21,22,29] has collected metadata on data use for restricted collections of papers, there is no general repository, no standardisation and no aggregation of results.

In this paper we propose that the community of information researchers create an infrastructure for the crowd-sourced collection of *metadata* about data use in the research

literature. Further, that such a collection should be licensed in an open manner and freely available for contribution and analysis.

Metadata about *papers* (such as authorship, affiliation and citation) is well-established as a valuable resource for understanding scholarly communication. Unfortunately there is far less metadata about the use of data (and code). Data citations are less formalized in the scientific literature than citations to papers and so are (currently) not amenable to automatic extraction. The proposal outlined here uses a crowdsourcing approach to address the lack of metadata about the use of data (and code). Citation databases allow a quantitative approach to understanding relationships between research papers; our current knowledge of data use is largely qualitative. Research about data use needs to be based on knowledge of current practice and we currently have little of the necessary metadata.

Quantitative measures of data use would enable the assessment of archiving policies of institutions, funding organisations, journals and individuals. Just as citation databases enable the calculation of summary measures (e.g. the h and g-index), metadata on data use would enable the creation of a data archiving index.

2. BACKGROUND

The history of scientific work can be seen as a collaborative endeavour spanning continents and centuries. Although involving numerous exceptions and contradictory sub-trends, one trend is the importance of building on prior work, and therefore the importance of knowing what that prior work is. From the proprietary secrecy of the alchemist we can see the growth of both formal and informal ways of scientists sharing findings, theories, methods and instruments. Citing prior work has evolved from a courtesy to an ethical necessity. Nowadays most non-commercial funding mandates publication, as do the processes for hiring and promotion. In recent years, accompanying rampant inflation in journal subscription rates, there has been a push for open access to scientific publications, particularly where that work has involved public funding.

Open access has been focused on the works themselves; the article in a journal. In parallel with the open access movement has been a growing awareness that much research is not easily confined to the boundaries of a largely textual article. The eScience and eResearch paradigms recognise that much research is backed by datasets and code that are important to the understanding, impact and reproducibility of the work. The issue of access to data, such as that provided in Supplemental Information to articles, is sometimes described as Open Science or Open Data [20]. Both open access and open data approaches share the sentiments expressed by Willensky:

> *A commitment to the value and quality of research carries with it a responsibility to extend the circulation of such work as far as possible and ideally to all who are interested in it and all who might profit by it*. (p. xii) [30]

Several arguments are made for greater open access, including equity and social justice, and that it will facilitate the progress of science as a whole [2,29]. Wicherts et al. simply state that "scientific evidence should be publicly accessible as a matter of principle" [29].

Although the advocacy of open data can be seen as a natural broadening of the arguments for open access, enabling open data is more challenging. The necessity of publication of scientific research has now been well established for centuries. The only argument is about how people should be able to access those published papers and how much, if anything, they should have to pay. By contrast, whether the data used in a paper should be published at all is still somewhat up for debate. And if a given community agrees that data should or must be shared, how is it to be done? There are a variety of approaches including informal exchange agreements between scientists [3,27], provision upon request, hosting data on a laboratory or university server, a disciplinary data repository, or as associated materials in a journal publisher's digital library [24]. We hope that more systematic methods will evolve over time, just as more systematic methods of publishing journal articles and citing prior work have evolved. Consequently, we believe a good starting point is to work towards a record of what data is actually available, where, and how to access it.

The Science Citation Index (SCI) was created by Eugene Garfield in 1960. Its development enabled a variety of uses including providing an alternative to subject indexing for accessing related work in other disciplines. It enabled the field of bibliometrics and various ways to automate the analysis of the impact of particular papers, journals and scientists.

> *While, at the time of the project's completion, the government sponsors chose not to subsidize the development of a national citation database, Eugene Garfield was encouraged to move ahead with the private publication of his multidisciplinary citation index as the first edition of the Science Citation Index® (SCI®). Available for purchase since 1963, the SCI then and now represents the most comprehensive citation index to the scientific journal literature. Today, the Web-based version of that index covers 5,600 journals across more than 150 scientific disciplines.* [28]

As this quotation from the current corporate owners of SCI shows, the citation information quickly moved into the private sphere, somewhat constraining repurposing of the data by the scientific community. The advantage of commercial ownership of such data is that it does not have to be initiated and continually directly supported by recurring public grants or be adopted by a foundation or library. Rather, public and private entities support it indirectly through purchases and subscriptions. Revenue enhancing innovations by the owners may lead to new uses and lower prices. The disadvantage is that a public corporation has a duty to maximize profits for its shareholders. Consequently there is a risk of a monopoly supplier of an information resource exploiting that power, as we have seen with journal prices. Also there is a strong temptation for the company to take a rather conservative approach to researchers requesting to do innovative things with their intellectual property (IP). There is always the risk of damaging profitability through unintended consequences, and little downside in inhibiting innovation with the use of the resource by those outside the company, who naturally have different interests than profit maximization. That is, saying yes to a request to use the data in an unorthodox way involves a hard to quantify risk, albeit tiny, of degrading the value of the IP in that data and thereby future profitability. Saying no just risks frustrating that researcher. Consequently there is a bias to conservatism with respect to data repurposing.

Similar concerns could be raised about such datasets under government or university control where access is not open so that permission must be sought. Government departments may have a mandate to generate revenues to save on taxes, so just like private companies may be reluctant to lose IP or risk revenue streams. Likewise a university hosting a dataset where permission must be

asked for unconventional use may worry about lost control, potential revenue foregone or embarrassing findings ensuing, again leading to a degree of conservatism.

2.1 Open Bibliographic Data

In contrast to large commercial databases such as the SCI there are several attempts to create bodies of open bibliographic data of the published research record [13]. However these also show both the strengths and weaknesses of public and open approaches to bibliographic data.

The RePEc (Research Papers in Economics) project illustrates the potential viability of open bibliographic data [15]:

> *the data input, i.e. the collection of bibliographic data about over 700,000 items of research (and growing by about ten thousands a month) is done by a large network of local volunteers, graduate students, faculty, secretaries, or IT professionals, who just follow a simple framework to organize their bibliographic data. Their individual cost in doing so is small, but once they realize their benefits in the circulation of their works, they are ready to do so.*

Krichel and Zimmermann argue that researchers, and institutions, have incentives to participate in "open academic libraries", such as RePEc, as they have an interest in accurate bibliographic data. They also note that:

> *It is only when authors and institutions are documented that they really start to make serious efforts in participating* [15]

RePEc deserves close scrutiny as a successful project in open metadata. Such analysis needs to be sociotechnical – we must look not just at the design of the resource, its interface and ease of use, but also the policies of entry, correction and use, as well as how the user community as a whole adopted and embraced the concept.

A detailed sociotechnical analysis would include amongst other things an examination of the differing incentives of various stakeholders to participate. This in turn requires studying their respective cost-benefit trade-offs in both the short and the long term. It is worth noting that the longer quotation above explicitly raises cost-benefit incentives – perhaps unsurprisingly for a resource aimed at economists. Time and space preclude a detailed sociotechnical analysis in this paper, but as a starting point for future analysis and discussion, we can note some factors in the design that are likely to have contributed to the success of the project. RePEc is a decentralized database of working papers, journal articles and software components. It integrates with existing university repositories through nightly updates, so it does not interfere with current practices and minimizes the amount of effort required to participate - chiefly in preparing and maintaining metadata describing publications. It provides a number of ways to use the information provided including an author service that is of use in maintaining a public academic presence, statistics on citations and downloads, and rankings. All RePEc information is freely available, facilitating the development of new uses and services. A focus on just economics research allows for growth through peer recommendation and an easier way to reach a critical mass.

By contrast, getCITED [12] appears to have enjoyed substantial initial success and growth for a number of years, but does not seem to have continued that early promise. Again we can't ascribe causality, but we can note suspicions. It aimed to cover all research, and so may not have been able to enjoy local-area network effects of a sub-discipline. This lack of focus may also affect the kind of publicity necessary to encourage adoption. We are unable to find information about data licensing, and easy re-use of data provided. Consequently the immediate incentives for participation may be lower. Of course if getCITED got to be large enough, then many network effect benefits would kick in. But what kinds of benefits motivate participation before critical mass is reached?

Other projects that use bibliographic data include DBLP, CiteULike, Mendeley, Zotero and LibraryThing. These mostly operate by harvesting extant bibliographic data (with varying amounts of information extraction, sophistication and reformatting). They still have the challenge of adoption – persuading people to bother to use them, particularly as the number of rival applications increase. This is addressed by providing various value-add features so that the effort of use is offset by the benefits provided by easy ways of collecting, organizing, commenting on and sharing sets of references. With greater participation, network effects can be exploited, including recommendations based on the selections and actions of others and the sharing of new or corrected bibliographic information manually entered by an individual. But as in RePEC, these network effects might best be considered as bonus features – even before they had a chance to kick in, it seems there was sufficient incentive to participate from the other features. It is worth noting that CiteULike [5] and DBLP [6] both allow downloads of data for further analysis by researchers; DBLP has itself become a data source for other researchers (e.g. [32])

The campaign for open access to journal articles has led to the development of SHERPA/RoMEO [25]: a searchable database of publishers' policies regarding the self- archiving of journal articles on the web and in Open Access repositories. This makes it easy to tell whether a given journal permits self-archiving. This does not guarantee that open access to an article in the journal actually exists, just that it is allowed. This serves several purposes. It lets an author know if she can make her paper freely available, for example in an institutional repository. It also has some similarities to what we are proposing for information about open data. It is a single resource that can be used to check a large proportion of all publishers' access policies. This allows the accumulation of aggregate statistics and temporal analyses of trends towards more open access, or at least trends of removal of barriers to the same.

SHERPA/RoMEO can also serve as a public inducement for publishers to adopt more liberal policies by the example of other publishers and pressure from their authors citing precedent from rival publication venues. The data it contains is public and available for further processing under a Creative Commons Attribution-NonCommercial-ShareAlike license. But unlike our proposal it is not collaboratively developed and curated. It is also much smaller in scope. Even if it expanded to cover all significant scientific journal publishers, this would still be far smaller than a listing of all publicly accessible datasets. An associated project, JULIET [14], includes information about certain funding organizations' data archiving policies. The information in JULIET is updated by community contributions using an online form to collect basic structured data or email.

In summary, citation data is largely concentrated in one commercial supplier, although some open bibliographic data systems have succeeded in particular niches.

2.2 Data Sharing and Archiving

We now move from open bibliographic citation data and information about open access to information about open data. At present metadata on data archiving is collected in individual studies of small sections of the published literature. The data from these studies is not aggregated and, as there are only a few such studies, then meaningful comparisons between disciplines are not yet feasible.

Piwowar has studied the availability of supporting data for published papers in the field of "biological gene expression microarray intensity values" [21]. Her study used 11,603 articles and applied automated methods for detecting associated datasets. The imprecision of automated evaluation led to an estimate that about 45% of the articles had provided supporting online data. Piwowar notes that:

> *It is disheartening to discover that human and cancer studies have particularly low rates of data sharing. These data are surely some of the most valuable for reuse, to confirm, refute, inform and advance bench-to-bedside translational research*

and concludes that:

> *the results presented here argue for action. Even in a field with mature policies, repositories and standards, research data sharing levels are low and increasing only slowly, and data is least available in areas where it could make the biggest impact*. [21]

Anderson et al. summarise several studies on data sharing in economics that found low rates of data availability, even in journals with specific data archiving policies [1]. Their work is a small example of the comparisons and trends that can be analysed when appropriate metadata is available. Wicherts et al. attempted to locate data from 249 studies in American Psychological Association journals, after extensive communication with authors they achieved a 26% success rate [29]. Savage and Vickers ran a similar study on a small sample of articles from two journals from the Public Library of Science (PLoS) [22]. One out of 10 authors "complied":

> *In conclusion, our findings suggest that explicit journal policies requiring data sharing do not lead to authors making their data sets available to independent investigators.* [22]

The small number of empirical studies on data availability have found that most data supporting published results is not available; however it is difficult to generalise across disciplines as there is a lack of metadata. These results are consistent with qualitative studies in which researchers give a variety of reasons for not sharing their data [3,27].

It is important to clarify the distinction between data citations and data availability. A data citation in a paper notes that a particular named dataset was used in the work reported in the paper. The citation may be in the references section, in a footnote or in the narrative of say the results section. Just as a traditional citation of a paper needs to provide sufficient information to unambiguously identify the paper being referred to, so must a data citation do for a dataset.

A data citation tells us a dataset has been used and which one; it does not mean anyone else can access it. Data availability tells us whether it is available and if so how and where to get it. A citation of a paper tells us the name and details of the work cited, but not whether it is freely available. Even if only available for payment, such a cited article is typically easier to access and verify than another kind of referred-to piece of information such as an opinion or quotation of another researcher noted in the article merely as "personal communication". Sadly, many datasets are equally inaccessible for verification, existing as "personal data".

Open data allows others to make use of data obtained with great effort, permitting combination, aggregation and repurposing. The history of science is full of examples of unexpected novel uses for prior work in different domains. In addition, open data makes it easier for others to check the validity of the findings and derived results reported in a paper, and the replication (or refutation) of the study in other settings. This allows for the identification and correction of errors, thereby ensuring greater reliability – especially necessary when the results of one piece of work rely on the veracity of others. This need is critical to the whole of science but becomes particularly acute when scientific findings (often including early and necessarily incomplete ones) are used to inform public debate and public policy.

Several authors have noted the connection between policies on data sharing and the wider implications for government policy and legislation. Anderson et al. provide a specific example:

> *Suppose, for example, that McCrary (2002) had not found the programming error reversing Levitt's (1997) result in the American Economic Review that increases in police substantially reduce crime. If a policymaker had acted on Levitt's finding and shifted funds from, say, low-income housing to police, social welfare would have been reduced*. [1]

McCullough and McKitrick describe several data-centred incidents from a variety of disciplines and suggest that greater diligence is needed when using data-backed research results for public policy [17]. The 'climate sceptic' blogosphere focuses on data availability issues precisely because of the policy implications of climate change research. An approach that targets the availability of data for policy-relevant research could be an effective campaign tool. Data sharing statistics have the potential to become a political tool: 'why are you implementing a policy which is based on un-reproducible research? Only 20% of the data supporting those research conclusions is actually available?'

In summary, the small number of studies on data sharing suggests that most data is not shared. However, information researchers lack the metadata to be able to make broad quantitative statements across diverse fields of research. This inability to make informed statements hinders our contribution to wider debates on the intersection of information use and public policy.

3. OPEN METADATA ON DATA USE

The previous section has made the case for the desirability of a resource containing metadata on data use, and particularly on which data is open. But how is that to be achieved? Just because something is desirable does not mean it is feasible. It can sound a noble but rather utopian idea, so how might we design something that has a realistic chance of success? We have looked at various projects that are analogous and whose sociotechnical designs can serve as useful inspirations (or warnings) in designing such a resource.

We propose that information researchers should facilitate the creation of metadata on data use by creating an infrastructure for crowd-sourcing: a data-use Metadatapedia.

Such a system would resemble Wikipedia but have a completely different notability criterion for inclusion: each entry would represent a single research paper. The intention is that each entry would move beyond the basic bibliographic information (as in getCited.org) to include structured data on the datasets used in that paper and their archiving status. Citations to other papers would not be significant however citations to data sets (or papers that represent data sets: 'data papers') would be included. Whether to start with a structured metadata template or to allow structure to evolve (as in the info-boxes in Wikipedia) is an open question.

We do not have a precise design to propose. Instead we want to make the case for a need and a set of approaches to developing such a design as a case study of sociotechnical design engineering. Many collective resources have failed, but a few have succeeded. What can we learn from these to increase the odds of success of a particular design? Much depends on the resultant cost-benefit trade-offs for different stakeholders over different timescales. It is no use noting that if everyone just pitches in, then network effects guarantee that in the end almost everyone will benefit to a far greater extent than their initial efforts. Rather there need to be immediate visible short term benefits from participating and contributing, with the network effects treated as a long term pure bonus once a critical mass has been reached. Understanding the incentives and disincentives to participate in data sharing [4] are important in designing a system that aims to adjust their impact. Equally, the incentives and disincentives to participate in Metadatapedia can be articulated and studied in use, and the design of the system adjusted appropriately. It is likely that designing features that provide direct benefits from collecting and sharing datasets of use or interest to an author will be a component of this, perhaps inspired by such features in some of the systems reviewed above.

There is a lot of information that may be useful to include about a given dataset. However the more that is required, or the more complex the options seem to be, the more daunting initial participation will seem to be. So in the interests of keeping adoption costs low, we need to consider what is the least amount of metadata necessary to be useful. The Wikipedia approach is one extreme case – anyone can start an article with just a few words and then hope others will be inspired to add to it. Wikipedia also makes it easier to provide partial and incremental information and to enable error checking and corrections. This is supported not just by the functionality and design of the wiki interface but also the processes, policies and norms of the Wikipedia community. However there are alternate approaches to seeding entries, such as the way that DBLP uses metadata from the ACM.

For many types of research papers the metadata needed would be as simple as one bit of information: this paper does not use any data sets. Significant sections of the scholarly record could probably be automatically marked as not using any data: for example, any item identified as a book review is likely to be non-data-using with a high degree of accuracy. Wider automatic approximations can be envisaged, any item in a philosophy journal could be tentatively marked as not using any datasets.

Although approximations can contribute, the main goal is for the infrastructure to facilitate human crowd-sourced information. This would include metadata on:

- Locations: such as URLs
- Identifiers: such as DOIs and accession numbers
- Licenses
- Formats

It is likely that beginning with a focus on one particular research domain and community will increase the odds of attaining a local critical mass that can be used to make a case and support organic growth first within that field and then in related fields. But which community should that be? The answer is likely to be opportunistic, depending on possible funding and personal interests and networks. But we can describe some characteristics of a pilot community that seem likely to increase the chances that the project will thrive while avoiding being such a special case that lessons learned here will not easily generalize:

- A very small sub-community with a strong sense of identity
- Part of a larger community and with clear links to other communities to facilitate future growth
- Having a clear need to share data
- At least some key members interested in open data
- Senior support including from major funding bodies and journals in the field
- Having some sceptics about open data needing to be convinced
- Distributed geographically with a significant international component
- Not focussed on truly enormous datasets (a separate problem entirely)
- Not using personally identifiable data or data derived in part from proprietary sources
- Dealing with somewhat diverse datasets so as to avoid being a special case with clear local solutions
- Tolerant of the inevitable errors that will occur in a pilot project
- Sufficiently resource-rich to be able to participate (most likely by having some students interested in the idea)

Given the size of the problem, relying on authors and data owners to be the sole information providers seems high risk. Various problems with faculty participation in institutional repositories show that despite numerous clear personal incentives, authors can be reluctant to participate**.** As a result a more open Wikipedia-like approach creates the opportunity for a wider constituency to participate. But who will bother and why? Is it possible to create incentives for others to participate?

We might imagine scientists (especially new scientists and students) assembling a personal collection of public datasets they find useful to consult, replicate, build upon, or use as practice data in their own work. By facilitating this we create an incentive to create metadata not just for one's own datasets but for those of others, even if the initial motivation is solely personal benefit from public data. This need not be too different from the practice of current students assembling sharable sets of citations of papers they need to read and refer to.

There are also educational uses that could be explicitly supported. It may be productive to develop meaningful learning activities for students of data curation involving participating in the retrospective recording and analysis of public datasets.

Metadatapedia would enable us to address the allegation from the introduction about the archiving history of individual researchers. More generally we can imagine formalising results into a D-

index: a data archiving index. Just as h and g indices assess citations, then the D-index reflects the archiving of supporting data and a related C-index, calculated in the same way, could reflect the archiving of supporting code. Whereas h and g indices are used to indicate the influence of researchers' work through citation, a D-index would provide an indicator of a separate facet of research activity: data archiving. A likely consequence of the creation of a reliable D-index would be the emergence of rankings and comparison tables. Just as citation-based indices have been used as inputs for rankings of researchers, institutions and journals, we might expect to see similar uses: which journal has the best data archiving performance from its authors? Although measures and rankings can easily become pernicious, there is the opportunity that ongoing study and reporting of results from the project can be used to encourage greater participation both in making data open, and in documenting that in Metadatapedia. How to do that appropriately becomes another design challenge of the project.

One study that a D-index would render feasible would be to investigate whether data sharing behaviours are connected to paper retractions: in other words, is there a relationship between a D-index and the proposed Retraction Index [9]? In general, we would expect that open re-usable metadata will enable a greater volume of research to be undertaken. In particular, we believe that an open environment will increase the ability of information research to provide relevant evaluation of data sharing policies such as the NSF Data Management Plan. We believe that communities such as the 'climate sceptics' will increasingly look to access the primary research data that underlies public policy, and that there will be a widespread need to accurately assess data archiving behaviour across many disciplines.

We propose this system now, at the start of large scale work on data archiving, to attempt to ensure the open nature of metadata on data use. A crude simplification would be to ask whether this metadata will end up like the citation data (in a restricted commercial environment) or like Wikipedia (in an open environment that facilitates re-use and research).

However, the open route will not happen automatically; indeed it may not happen at all. Specifically, we suggest that it is unlikely to happen on its own and that information researchers should take actions to preserve the potential open-nature of this type of metadata.

The key features of our proposed system are:

- Machine readable data
- Clear open licensing framework that facilitates re-use and the development of novel applications and benefits from the available data
- Seed-able from other open sources of metadata
- Amenable to computer-derived data additions (e.g. the 'philosophy approximation', automated studies [21])
- A level of simplicity on a par with Wikipedia, to reduce the costs of participation
- Ease of reconfiguration of data structures, features, processes and norms, inspired by the historic adaptability of these in Wikipedia
- A design focus on creating immediate benefits from participation, not relying on the longer term benefits from critical mass
- The determination of a particular research community in order to create an existence proof of the benefits arising from attaining local critical mass

A desirable feature would be the support of organisations of information researchers such as ASIS&T and the iSchools Caucus.

4. CONCLUSION

If we, as a community of information researchers, wish to understand data archiving across the academic literature then we need to access the appropriate metadata. This metadata could be restricted, like citation data, or it could be open, like Wikipedia. One possible route to an open environment is to create an enabling infrastructure to both crowd-source the metadata and aggregate our own research results.

There is no guarantee that such a proposal will be a success: crowd-sourcing does not automatically equal success [8]. However, if the alternatives are either a commercially owned, and therefore controlled, metadata collection (or no collection at all) then we should explore the open route. The proposed approach outlined here, even if not practical in its current form, can serve as a useful initiating thought experiment to inform new and better design ideas. The studies that would be enabled through open metadata on data use are likely to be informative both about the nature of research itself and its relevance to public policy.

5. REFERENCES

[1] Anderson, R.G., W.H. Greene, B.D. McCullough and Vinod, H.D. 2008. The Role of Data/Code Archives in the Future of Economic Research. *Journal of Economic Methodology*. 15, 1, 99-119.

[2] Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P. 2004. Promoting access to public research data for scientific, economic, and social development, *Data Science Journal*, 3, 135-152.

[3] Borgman, C.L. 2007. *Scholarship in the Digital Age: information, infrastructure, and the internet*. Cambridge, MA: MIT Press.

[4] Costello, M.J. 2009. Motivating online publication of data, *BioScience*, 59, 5, 418–427.

[5] CiteULike. http://www.citeulike.org/

[6] DBLP Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/db/

[7] Delserone, L.M. 2008. At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries. *Library Trends* 57, 2, 202-210.

[8] Eickhoff, C. and de Vries, A. 2011. How Crowdsourcable is Your Task? *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM),* 11–14.

[9] Fang, F.C. and Casadevall, A. 2011. Retracted Science and the Retraction Index, *Infection and Immunity*, 79, 10, 855-3859.

[10] Freese, J. 2007. Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods & Research*, 36, 2, 153-172.

[11] Fry, J., Schroeder, R. and den Besten, M. 2008. Open science in e-science: contingency or policy? *Journal of Documentation*, 65, 1, 6-32.

[12] getCITED. http://www.getcited.org/

[13] JISC. Open Bibliographic Data Guide. http://obd.jisc.ac.uk/

[14] JULIET http://www.sherpa.ac.uk/juliet/index.php

[15] Krichel, T. and Zimmermann, C. 2009. The Economics of Open Bibliographic Data Provision. *Economic Analysis and Policy* 39, 1, 143-152.

[16] Laine C., Goodman S.N., Griswold M.E. and Sox, H.C. 2007. Reproducible research: moving toward research the public can really trust. *Annals of Internal Medicine*, 146, 6, 450-453.

[17] McCullough, B.D. and McKitrick, R. 2009. *Check the Numbers: The Case for Due Diligence in Policy Formation*, Studies in Risk & Regulation. Fraser Institute. February 2009.

[18] McIntyre, S. 2006. Comment at Climate Audit, 24 July 2006. http://climateaudit.org/2006/07/06/new-thompson-article-at-pnas/#comment-55284

[19] Mervis, J. 2010. NSF to Ask Every Grant Applicant for Data Management Plan, *Science Insider*, 5 May. http://news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html

[20] Murray-Rust, P. 2008. Open data in science. *Serials Review* 34, 1, 52-64.

[21] Piwowar, H.A. 2011. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS ONE* 6, 7, e18657. doi:10.1371/journal.pone.0018657

[22] Savage C.J. and Vickers A.J. 2009. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE* 4, 9, e7078. doi:10.1371/journal.pone.0007078

[23] Science Foundation Ireland, 2010. *Policy Relating to the Open Access Availability of Published Research*. http://www.sfi.ie/assets/files/downloads/Funding/grant_policies/open%20access%20dec%2010.pdf

[24] Science Staff 2011. Introduction to Special Issue Challenges and Opportunities, *Science*, 331, 6018, 692-693.

[25] SHERPA/RoMEO: Publisher copyright policies & self-archiving. http://www.sherpa.ac.uk/romeo/

[26] Stodden, V. 2010. Open science: policy implications for the evolving phenomenon of user-led scientific innovation. *Journal of Science Communication*, 9, 1. http://jcom.sissa.it/archive/09/01/Jcom0901(2010)A05

[27] Tenopir C., Allard S., Douglass K., Aydinoglu A.U., Wu L., Read, E., Manoff, M. and Frame, M. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE,* 6, 6, e21101. doi:10.1371/journal.pone.0021101

[28] Thomson Reuters, *History of Scientific Indexing*. http://thomsonreuters.com/products_services/science/free/essays/history_of_citation_indexing/

[29] Wicherts, J.M., Borsboom, D., Kats, J. and Molenaar, D. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist*, 61, 726-728.

[30] Willinsky, J., 2006. *The Access Principle: The case for open access to research and scholarship.* Cambridge, MA: MIT Press.

[31] Xia, J. 2007. Assessment of self-archiving in institutional repositories: across-disciplines, *Journal of Academic Librarianship*, 33, 6, 647-654.

[32] Zaïane, O.R., Chen, J., and Goebel, R. 2007. DBconnect: mining research community on DBLP data. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*. ACM, New York, NY, 74-81.

Cite as:

Nichols, D.M., Twidale, M.B. and Cunningham, S.J. (2012) Metadatapedia: a proposal for aggregating metadata on data archiving. *Proceedings of the 2012 iConference (iConference '12).* ACM, New York, NY, USA, 370-376.

http://doi.acm.org/10.1145/2132176.2132224